On Power-Law Distributed Balls in Bins and its Applications to View Size Estimation

Ioannis Atsonios, Olivier Beaumont, Nicolas Hanusse, Yusik Kim

On Power-Law Distributed Balls in Bins and its Applications to View Size Estimation

Ioannis Atsonios, Olivier Beaumont, Nicolas Hanusse and Yusik Kim

CNRS and INRIA Bordeaux – Sud-Ouest, University of Bordeaux, LaBRI, France

Abstract. The view size estimation plays an important role in query optimization. It has been observed that many data follow a power law distribution. In this paper, we consider the balls in bins problem where we place balls into N bins when the bin selection probabilities follow a power law distribution. As a generalization to the coupon collector's problem, we address the problem of determining the expected number of balls that need to be thrown in order to have at least one ball in each of the N bins. We prove that $\Theta(\frac{N^\alpha \ln N}{c_N^\alpha})$ balls are needed to achieve this where α is the parameter of the power law distribution and $c_N^\alpha = \frac{\alpha-1}{\alpha-N^{\alpha-1}}$ for $\alpha \neq 1$ and $c_N^\alpha = \frac{1}{\ln N}$ for $\alpha = 1$. Next, when fixing the number of balls that are thrown to T, we provide closed form upper and lower bounds on the expected number of bins that have at least one occupant. For n large and $\alpha > 1$, we prove that our bounds are tight up to a constant factor of $\left(\frac{\alpha}{\alpha-1}\right)^{1-\frac{1}{\alpha}} \leq e^{1/e} \simeq 1.4$.

1 Introduction

1.1 Context

Datacubes and Query Optimizations Query optimization can be decomposed into several steps. One of the most important deals with the estimation of the memory and time requirements of some possible sequences of operations in order to choose the cheapest. In the particular case of OLAP queries, the use of a data structure called a *datacube* [GBLP96] allows to speed up the queries. It consists of the storage of some *views* corresponding to intermediate results. Usually, each materialized view is a cuboid, that is the set of aggregative values populating a fact table for a given combination of attributes. View selection algorithms rely on the fast estimation of the cuboids.

The simplest way to estimate a view size is to scan the whole dataset. In practice, scanning once a dataset of a few millions of entries can take 1 minute. For a d-dimensional data set of large size, the computation of the exact size of the 2^d views can take up to a few days. Whenever data are dynamic, this is not realistic. A quick but naive way to estimate the size of a view v is to use Cardenas formula [Car75],

$$\mathbf{E}[\text{viewsize}] = m(v)\left(1 - \left(1 - \frac{1}{m(v)}\right)^T\right). \quad (1)$$

where $m(v)$ denotes the number of all possible value combinations of view v, and $T = |\mathcal{T}|$, the number of entries in fact table $\mathcal{T}$. This formula holds true if the entries of $\mathcal{T}$ are chosen uniformly at random from the set of all possible attribute combinations. However, it turns out that this estimate leads to overestimate the real size as soon as the uniformity assumption does not hold. From this observation, several more sophisticated approaches have been proposed (see [AL07] for an experimental comparison). For instance, it has been proposed in [CPKH05,DF03,GT01,FM85] (for a data stream setting [KNW10]) to scan $\mathcal{T}$ but with a light consumption of memory, without any assumption on the distribution of data. These algorithms are based on independent and uniform hashing and provide a theoretical accuracy of $1 + \Theta(1/\sqrt{M})$. However, in practice, getting independent and uniform hashing is not realistic. In [AL07], it is shown that using only few kbytes is enough to get a good accuracy. When $\mathcal{T}$ can be partitioned, distributed computation of such estimates has been proposed in [BGH$^+$09], but the time complexity remains proportional to T. Nevertheless, whenever the number of view size requests becomes too large, this approach cannot be considered.

A second approach is based on sampling. In this context, the fact table is sampled and the skew of the data distribution is approximated by a statistical model. It turns out that without any assumption on the distribution of data, it is impossible to get an estimate with a good accuracy using a small sample size [CCMN00]. For instance, Faloutsos *et al.* [FMS96] choose a multifractal model based only on the knowledge of the number of occurrences of the most frequent tuples in the sample $\mathcal{T}'$ and $|\mathcal{T}'|$. Haas *et al.* [HNSS95] propose an estimator based on the distribution of data in the sample. Some work [NT03,AL07] report the relevance of the multifractal model with respect to the accuracy but there is no theoretical guarantee on the accuracy related to the size of the sample. We point out that in [NT03], an estimator dedicated to Pareto distribution (similar to power law for $\alpha > 1$) is proposed. In this approach, time and memory space are proportional to $|\mathcal{T}'|$. Motwani and Vassilvitskii [MV06] propose a sampling based method under the power law assumption that provides near accurate results with positive probability. More precisely, denoting by $\mathbf{F}_\alpha(T, n)$ the expected number of filled bins after T trials over n bins, and assuming that the exponent of the power law distribution α is known, they propose an algorithm providing an estimator of $\mathbf{F}_\alpha(T, n)$ with an accuracy of $(1 + \epsilon)$ with probability $2\exp(-\epsilon^2\mathbf{F}_\alpha(T, n))$, using a sample of size $\Theta((1 + \epsilon)^{1+\alpha}\mathbf{F}_\alpha(T, n)c_n^\alpha)$ with c_n^α being a normalizing factor.

Boneh and Hofri [AM97] provide the distribution of the number of filled bins for general bin selection distributions and derive the expected value (See Equation 3), which, in this form, is intractable to compute due to $|\mathcal{T}|$ being excessively large. To summarize, to our knowledge, no existing estimate provides at low computational cost an accurate estimate of $\mathbf{F}_\alpha(T, n)$ for power law distributed data, even if power law parameter α is given.

1.2 Model

We abstract the problem of determining the view size as a *balls in bins* problem. The setting of the balls in bins problem is that there are some empty bins where balls are thrown into. Balls decide, independently of each other, which bin to fall into according to some given probability distribution over the bins. Among the many variants of problems arising from this setting, we are particularly interested in the following two questions. Given a fixed number of bins and bin selection probabilities, what is the expected number of balls that need to be thrown in order to have at least 1 ball in all bins? (equivalent to the coupon collector's problem) *and* Given a fixed number of bins, a fixed number of balls, and bin selection probabilities, what is the expected number of bins that have at least 1 ball in it?

The relation between the view size of a fact table and the balls in bins problem is based on the following analogy. A **bin** corresponds to a possible combination of attribute values, i.e., a single cell of the cuboid and a **ball** corresponds to a single row of the fact table that contributes to the count of exactly one cell of the cuboid. Thus, the number of "occupied" bins corresponds to the view size. From a theoretical point of view, once the bin selection probabilities $P = \{p_1, p_2, \ldots, p_n\}$ are given, estimating a view size from a fact table of T entries can be modeled by a generalized version of balls and bins problem. At each time step t going from 1 to T, a ball is thrown in one of the bins, and bin i is chosen with probability p_i which follows a power law, *i.e.* $p_i = \frac{c_n^\alpha}{i^\alpha}$ where $\alpha \geq 0$ and c_n^α is the normalizing factor of the form $c_n^\alpha = \left[\sum_{i=1}^{n} \frac{1}{i^\alpha}\right]^{-1}$.

We focus on $\mathbf{F}_P(T, n)$, the expected number of filled bins after T trials, corresponding to the number of different values of a combination of attributes. For convenience, whenever P follows a power law distribution of parameter α, we use the notation $\mathbf{F}_\alpha(T, n)$. $\mathbf{F}_0(T, n)$ is a very well known uniform and $\mathbf{F}_P(T, n)$ is given by the Cardenas formula [Car75]. From the coupon collector problem [MR95], we know that whenever $T < n \ln n$, $\mathbf{F}_0(T, n) < n$ and if $T = n \ln n$, $\mathbf{F}_0(T, n) = n$ with constant probability. For $\alpha = 1$, T needs to be $n \ln_2 n$ in order to get $\mathbf{F}_1(T, n)$ with constant probability [BP96]. To the best of our knowledge, for $\alpha \notin \{0, 1\}$, no other closed formula or bounds on $F_\alpha(T, n)$ are known.

1.3 Our contribution

Expected time to fill all the bins The coupon collector is a well-known problem related to estimating the number of balls required to fill at least once every bin. In our setting and assuming that the frequencies of tuples follow a power law of parameter α, we raise the question on computing the expected number of entries $\mathbf{E}[T]$ required in order to make a view v saturated, that is of size $n = m(v)$. As an easily obtainable upper bound of $\mathbf{E}[T]$, we have $O(n^{\alpha+1} \log n)$. This can be derived by making bin i reject a ball with probability $(p_i - p_n)/p_i$. In this paper, we prove that (Theorem 1)

$$\frac{(n - \sqrt{n})^\alpha}{c_n^\alpha} \ln(\sqrt{n} + 1) \leq \mathbf{E}[T] \leq \frac{n^\alpha}{c_n^\alpha}(1 + \ln n)$$

$$\text{and } \lim_{n\to\infty} \frac{\mathbf{E}_n[T]}{n^\alpha(1+\ln n)/c_n^\alpha} = 1,$$

where $\mathbf{E}_n[T]$ denotes the expected time to fill all n bins (Theorem 2).

Expected number of bins filled when throwing T balls We provide bounds on $\mathbf{F}_\alpha(T,n)$. For upper bounds, we prove that

- When $\alpha \geq 0, \alpha \neq 1$, then $\mathbf{F}_\alpha(T,n) \leq \frac{\alpha}{\alpha-1}(c_{n,\alpha}T)^{1/\alpha} - \frac{c_{n,\alpha}T}{\alpha-1}n^{1-\alpha}$ (Th. 6)
- When $\alpha = 1$, then $\mathbf{F}_\alpha(T,n) \leq \frac{T}{\ln n+\gamma} - 1 + T\left(1 - \frac{\ln T - \ln(\ln n+\gamma)+\gamma}{\ln n+\gamma}\right)$ (Th. 6), where $\gamma \simeq 0.5772$ is the Euler-Mascheroni constant.

For lower bounds, we have results for the general case $n \geq 1, \alpha \geq 0$ and tighter bounds when n is sufficiently large

- When $n \geq 1$ and $\alpha \geq 0$, then $\mathbf{F}_\alpha(T,n) \geq \left(1 - \left(1 - \frac{1}{T}\right)^T\right)(c_n^\alpha T)^{1/\alpha}$ (Th. 5)
- When n is large enough and $\alpha > 1$, then $\mathbf{F}_\alpha(T,n) \geq (T+1)^{1/\alpha} - 1$ (Th. 3)
- For $n \geq 1$ and $0 \leq \alpha < 1$, $\mathbf{F}_\alpha(T,n) \geq (n+1)\left(1 - e^{-\frac{(1-\alpha)T}{n+1} - \ln\frac{n+1}{n}}\right)$ (Th. 4)

At last, when n and $\alpha > 1$, we prove that the ratio between the upper and lower bounds is bounded by $e^{1/e}$, i.e., $UB/LB \leq e^{1/e} \simeq 1.4447$ (Corollary 2).

2 Bounds for the expected number of balls needed to fill all bins

Consider having n bins where we throw balls into. Assume that bin i has a probability p_i of being chosen to put a ball into in each trial. We are interested in finding the expected number of balls (or equivalently, trials), which we denote by T, necessary to fill each of the n bins at least once when the probabilities p_i follow a power law distribution.

The classical coupon collector's problem is the special case of this problem when the probabilities p_i follow a uniform distribution instead of the power law. For this special case, the expected number of balls necessary to fill all bins is given by nH_n, where H_n is the harmonic number defined as $H_n = \sum_{k=1}^{n} \frac{1}{k}$. The simple form for the solution of the classical problem benefits from the simple state space only requiring the number of currently occupied bins, whereas in the general case (without the uniformity assumption), it is necessary to keep track of the specific combination of bins that are currently occupied when doing the calculation. For large n, this calculation is intractable, and therefore, we focus on finding good bounds for T.

2.1 Stochastic majorization scheme for finding bounds

The main idea of our method to find upper and lower bounds for T is to consider models which are different from our original model where it is possible to

"order" the expected number of necessary balls through stochastic majorization arguments. We provide basic definitions and properties of stochastic orders. See [MJ94] for details.

Definition 1. *X is greater than Y in the stochastic order, written as $X \geq_{st} Y$ if $F_X(x) \leq F_Y(x)$ for all x.*

An alternative characterization of stochastic order is that $X \geq_{st} Y$ if and only if there exist X' and Y' on the same probability space where $P(X' \geq Y') = 1$ and have the same distribution as X and Y, respectively. Therefore, we use a coupling argument as a method of proof. From the definition, it is straightforward to verify that stochastic order implies ordering of the mean. To establish an upper bound, we need to have a model that is more pessimistic than the original model in occupying an empty bin in each trial. Let us denote the original model by O and let us consider a new model, denoted model U, that consists in $n+1$ bins where $p_i = p_n$ for $i = 1, 2, ..., n$ and $p_0 = 1 - np_n$ representing a "trash" bin (See Figure 1 for an illustration). Comparing this with our original model, intuitively,

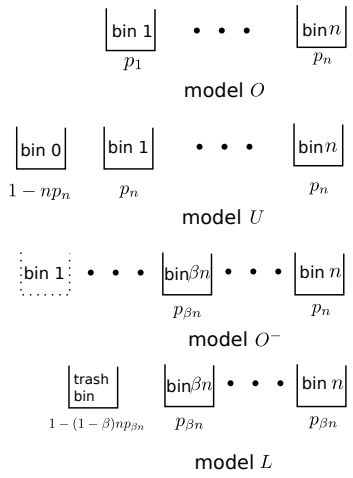

Fig. 1: Illustration of various models used. The probability of selecting a certain bin written below the bins.

balls that would normally go into a particular empty bin under model O will now have a non zero probability of going into the trash bin under the model U, failing to increment the number of occupied bins. Thus, the overall number of balls required to fill bins 1 through n under the model U will be in some sense larger than under the model O as formally stated in the following lemma.

Lemma 1. *Consider the random experiment of throwing balls independently into n bins with selection probabilities $p_1 \geq \cdots \geq p_n$ for the n bins. Call this experiment O and denote the random variable counting the number of balls thrown*

until there are no longer any empty bins by X. Now consider a random experiment of throwing balls independently into $n+1$ bins with selection probabilities $p_n, ..., p_n$ for the first n bins and $1-np_n$ for the last bin. Call this experiment U and denote the random variable counting the number of balls thrown until there are no longer any empty bins among the first n bins by X'. Then $X' \geq_{st} X$ and consequently, $\mathbf{E}[X'] \geq \mathbf{E}[X]$.

Proof. See Appendix A.1

Similarly, to find a lower bound, we need a model that is more optimistic. Consider a subset of the original model, which we will call model O^- that only includes the last $(1-\beta)n$ bins, i.e., $i = \beta n, ..., n$, where $0 \leq \beta \leq 1$. Suppose n is large enough so that the effect of rounding to an integer is negligible. Now consider an alternative model, which we will call model L, with $(1-\beta)n$ bins each having probability $p_{\beta n}$ of being chosen plus a trash bin having probability $1-(1-\beta)np_{\beta n}$ of being chosen. Since $p_{\beta n}$ is the largest choice probability among the bins in model O^-, loosely speaking, it will take less time to fill the $(1-\beta)n$ bins of model L than the last $(1-\beta)n$ bins of model O^-. Since filling the last $(1-\beta)n$ bins is a necessary condition for filling all n bins of the of the original problem, the number of balls required to fill all bins in model L provides a lower bound to the original problem (see Figure 1 for an illustration).

Lemma 2. *Consider the random experiment of throwing balls independently into $n-m+1$ bins with selection probabilities $p_m \geq \cdots \geq p_n$ and $\sum_{i=m}^{n} p_i \leq 1$. Note that it is possible that no bin is selected. Call this experiment O^- and denote the random variable counting the number of balls thrown until there are no longer any empty bins by X. Now consider a random experiment of throwing balls independently into $n-m+1$ bins with selection probabilities $p_m, ..., p_m$, provided $(n-m+1)p_m \leq 1$. Call this experiment L and denote the random variable counting the number of balls thrown until there are no longer any empty bins among the first n bins by X'. Then $X' \leq_{st} X$ and consequently, $\mathbf{E}[X'] \leq \mathbf{E}[X]$.*

Proof. omitted (same technique as Lemma 1)

To find $\mathbf{E}[X']$, we rely on the following lemma.

Lemma 3. *We are given n bins with equal probability of being filled with $p_i = p, \ \ i = 1, ..., n$ and $np \leq 1$. Then the expected time to fill all the bins is H_n/p.*

Proof. See Appendix A.2

We are now in a position to propose an upper and lower bound for $\mathbf{E}[T]$. For the upper and lower bounds, we apply Lemma 3 to models U and L respectively.

Corollary 1. *Under the power law assumption for the p_i,*

$$\frac{(\beta n)^\alpha H_{(1-\beta)n}}{c_n^\alpha} \leq \mathbf{E}[T] \leq \frac{n^\alpha H_n}{c_n^\alpha} \tag{2}$$

for any β that satisfies the conditions $(1-\beta)np_{\beta n} \leq 1$ and $0 \leq \beta \leq 1$.

2.2 General bounds when $n \geq 1$

In the view estimation context, n represents the number of different values an attribute can take. Thus, it is important to prove results that still hold true for smaller values of n. In this section, we provide bounds for $\mathbf{E}[T]$ when $n \geq 1$. This is achieved by finding bounds for H_n and applying Corollary 1.

Theorem 1. *For $\alpha \geq 0$ and $n \geq 1$ bins, the expected number of balls needed to fill all bins satisfies* $\frac{1}{c_n^\alpha} \cdot (n - \sqrt{n})^\alpha \ln(\sqrt{n} + 1) \leq \mathbf{E}[T] \leq \frac{n^\alpha}{c_n^\alpha}(1 + \ln n)$.

Proof. See Appendix A.3

2.3 Asymptotically accurate estimator for the expected number of balls

The following theorem proves that the expected number of balls needed to fill all bins when n is large.

Theorem 2. *Let $\mathbf{E}_n[T]$ denote the expected number of balls needed to fill all n bins. Then, for $\alpha \geq 0$,* $\lim_{n\to\infty} \frac{E_n[T]}{n^\alpha(1+\ln n)/c_n^\alpha} = 1$.

Proof. See Appendix A.4

3 Estimating the expected number of occupied bins given a fixed number of balls

In this section, we consider the problem of estimating the expected number of filled bins given that a fixed number of balls are thrown. Here, *filled bin* means that there is at least one ball in the bin. Let us assume that there are n bins having probabilities p_i for $i = 1, 2, ..., n$ and we throw T balls. To find an expression for the expected number of filled bins, let $Z_i = \mathbf{1}\{\text{bin } i \text{ is filled after } T \text{ throws}\}$. Then the number of filled bins at the end of T throws is $\sum_i Z_i$ and therefore the expected value of the number of filled bins after T throws is

$$\begin{aligned}\mathbf{E}\left[\sum_{i=1}^{n} Z_i\right] &= \sum_{i=1}^{n} \Pr\{\text{bin } i \text{ is filled after } T \text{ throws}\} \\ &= \sum_{i=1}^{n}(1 - \Pr\{\text{bin } i \text{ is not filled after } T \text{ throws}\}) = n - \sum_{i=1}^{n}(1 - p_i)^T\end{aligned} \tag{3}$$

When bin probabilities follow a uniform distribution, the expression reduces to a closed form without the summation. This yields to Cardenas' formula

$$n - n\left(1 - \frac{1}{n}\right)^T. \tag{4}$$

However, in the general case and even under our power law distribution assumption on the bin probabilities, it is difficult to express it without the summation. Moreover, in the context of computing the view size for database queries, n can be as large as 10^{20} in some cases, which makes the summation intractable. So we seek upper and lower bounds to our quantity of interest.

3.1 Lower bound

To obtain a lower bound to the number of filled bins, we must find a scheme that is pessimistic in the sense that the probability that a new bin will be filled at each trial is minimized. So we consider the following process. After each throw, if the ball goes into a previously empty bin, the ball is relocated to the remaining empty bin with the highest probability. This is to ensure that the probability that a new bin will be occupied at the next trial is minimized. After T throws, the random variable that counts the number of filled bins for this scheme is stochastically smaller than that of the original problem and therefore the expected values are ordered accordingly. Under this modified process, let X_T denote the number of filled bins after T throws. Then $\{X_T\}$ is a Markov Chain with the following transition rule

$$X_{T+1} = \begin{cases} X_T & \text{with prob. } \sum_{i=1}^{X_T} \frac{c_n^\alpha}{i^\alpha} \\ X_T + 1 & \text{with prob. } \sum_{i=X_T+1}^{n} \frac{c_n^\alpha}{i^\alpha}. \end{cases}$$

Note that it is not computationally hard to evaluate c once α and n are given. However, we use the approximation

$$c_n^\alpha \simeq \left(1 + \int_1^n x^{-\alpha} dx\right)^{-1} = \frac{\alpha - 1}{\alpha - n^{1-\alpha}} \qquad \text{when } \alpha \neq 1 \tag{5}$$

$$= \frac{1}{\ln n + \gamma} \qquad \text{when } \alpha = 1, \tag{6}$$

where γ is the Euler-Mascheroni constant. In particular, when n is large and $\alpha > 1$, $c_n^\alpha \simeq (\alpha - 1)/\alpha$.

For large n and $\alpha > 1$ When the number of bins n is sufficiently large and $\alpha > 1$, we provide a lower bound for the expected number of filled bins when T balls are thrown.

Theorem 3. *Let Z be the number of filled bins when T balls are thrown into n bins. When n is large and $\alpha > 1$, $\mathbf{E}[Z] \geq (T+1)^{1/\alpha} - 1$.*

Proof. See Appendix B.1

For any $n \geq 1$ and $\alpha < 1$ For the case $\alpha < 1$, we do not need n to be large to obtain a lower bound for $\mathbf{F}_\alpha(T, n)$.

Theorem 4. *For $0 \leq \alpha < 1$, a lower bound on the expected number of filled bins when T balls are thrown is given by $(n+1)\left(1 - e^{-\frac{(1-\alpha)T}{n+1} - \ln \frac{n+1}{n}}\right) - 1$.*

Proof. See Appendix B.2

For any $n \geq 1$ and any $\alpha \geq 0$ Here we provide lower bounds for the most general case. Notably, it covers the case when $\alpha = 0$.

Theorem 5. *For $\alpha > 0$ and $n \geq 1$, a lower bound on the expected number of filled bins when T balls are thrown is given by $\left(1 - \left(1 - \frac{1}{T}\right)^T\right) (c_n^\alpha T)^{1/\alpha}$.*

Proof. See Appendix B.3

Note that when T is large, the lower bound can be approximated by $\left(1 - e^{-1}\right) (c_n^\alpha T)^{1/\alpha}$.

3.2 Upper bound

Let us provide an upper bound of $\mathbf{F}_\alpha(T, n)$ that holds for $n \geq 1$ and $\alpha \geq 0$.

Theorem 6. *For $n \geq 1$, an upper bound on the expected number of filled bins when T balls are thrown is given by*

$$\begin{cases} \min\{(4), (15)\} & \alpha = 1 \\ \min\{(4), (14)\} & \alpha \neq 1 \end{cases}$$

Proof. See Appendix B.4

3.3 Bound performance

Corollary 2. *When n is sufficiently large and $\alpha > 1$, the ratio between the lower bound of Theorem 3 and the upper bound of Theorem 6 is less than or equal to $\left(\frac{\alpha}{\alpha - 1}\right)^{1 - \frac{1}{\alpha}} \leq e^{1/e}$.*

Proof. See Appendix B.5

4 Conclusion

The power law is a distribution frequently observed in real data sets. For the balls into bins problem, we studied the special but important case where the bin selection probabilities follow a power law. Asymptotically accurate estimators for the expected number of balls needed to be thrown in order to have all bins occupied as well as closed form expressions for the lower and upper bounds for the expected number of bins occupied when throwing a fixed number of balls are provided.

References

AL07. Kamel Aouiche and Daniel Lemire. A comparison of five probabilistic view-size estimation techniques in olap. In Il-Yeol Song and Torben Bach Pedersen, editors, *DOLAP*, pages 17–24. ACM, 2007.

AM97. Arnon Boneh and Micha Hofri. The coupon-collector problem revisited - A survey of engineering problems and computational methods. *Stochastic Models*, 13(1):39–66, 1997.

BGH+09. Kevin Beyer, Rainer Gemulla, Peter J. Haas, Berthold Reinwald, and Yannis Sismanis. Distinct-value synopses for multiset operations. *Commun. ACM*, 52:87–95, October 2009.

BP96. Shahar Boneh and Vassilis G. Papanicolaou. General asymptotic estimates for the coupon collector problem. *J. Comput. Appl. Math.*, 67:277–289, March 1996.

Car75. Alfonso F. Cardenas. Analysis and performance of inverted data base structures. *Commun. ACM*, 18(5):253–263, 1975.

CCMN00. Moses Charikar, Surajit Chaudhuri, Rajeev Motwani, and Vivek R. Narasayya. Towards estimation error guarantees for distinct values. In *PODS*, pages 268–279. ACM, 2000.

CPKH05. Min Cai, Jianping Pan, Yu-Kwong Kwok, and Kai Hwang. Fast and accurate traffic matrix measurement using adaptive cardinality counting. In Subhabrata Sen, Chuanyi Ji, Debanjan Saha, and Joe McCloskey, editors, *MineNet*, pages 205–206. ACM, 2005.

DF03. Marianne Durand and Philippe Flajolet. Loglog counting of large cardinalities (extended abstract). In Giuseppe Di Battista and Uri Zwick, editors, *ESA*, volume 2832 of *Lecture Notes in Computer Science*, pages 605–617. Springer, 2003.

FM85. Philippe Flajolet and G. Nigel Martin. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.*, 31(2):182–209, 1985.

FMS96. Christos Faloutsos, Yossi Matias, and Abraham Silberschatz. Modeling skewed distribution using multifractals and the '80-20' law. In T. M. Vijayaraman, Alejandro P. Buchmann, C. Mohan, and Nandlal L. Sarda, editors, *VLDB'96, Proceedings of 22th International Conference on Very Large Data Bases, September 3-6, 1996, Mumbai (Bombay), India*, pages 307–317. Morgan Kaufmann, 1996.

GBLP96. Jim Gray, Adam Bosworth, Andrew Layman, and Hamid Pirahesh. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-total. In *ICDE*, pages 152–159, 1996.

GT01. Phillip B. Gibbons and Srikanta Tirthapura. Estimating simple functions on the union of data streams. In *SPAA*, pages 281–291, 2001.

HNSS95. Peter J. Haas, Jeffrey F. Naughton, S. Seshadri, and Lynne Stokes. Sampling-based estimation of the number of distinct values of an attribute. In *VLDB*, pages 311–322, 1995.

KNW10. Daniel M. Kane, Jelani Nelson, and David P. Woodruff. An optimal algorithm for the distinct elements problem. In Jan Paredaens and Dirk Van Gucht, editors, *PODS*, pages 41–52. ACM, 2010.

MJ94. Moshe Shaked and J. George Shanthikumar. *Stochastic orders and their applications*. Academic Press, 1994.

MR95. Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

MV06. Rajeev Motwani and Sergei Vassilvitskii. Distinct values estimators for power law distributions. In *2006 Proceedings of the Third Workshop on Analytic Algorithmics and Combinatorics*, 2006.

NT03. Thomas P. Nadeau and Toby J. Teorey. A pareto model for olap view size estimation. *Information Systems Frontiers*, 5(2):137–147, 2003.

A Proofs of Section 2

A.1 Proof of Lemma 1

Lemma 1 *Consider the random experiment of throwing balls independently into n bins with selection probabilities $p_1 \geq \cdots \geq p_n$ for the n bins. Call this experiment O and denote the random variable counting the number of balls thrown until there are no longer any empty bins by X. Now consider a random experiment of throwing balls independently into $n+1$ bins with selection probabilities $p_n, ..., p_n$ for the first n bins and $1 - np_n$ for the last bin. Call this experiment U and denote the random variable counting the number of balls thrown until there are no longer any empty bins among the first n bins by X'. Then $X' \geq_{st} X$ and consequently, $\mathbf{E}[X'] \geq \mathbf{E}[X]$.*

Proof. Let

$$Z_{jk} = \mathbf{1}\{\text{trial } k \text{ selects bin } j \text{ under experiment } O\}$$
$$Z'_{jk} = \begin{cases} Z_{jk} & \text{with probability } p_n/p_j \\ 0 & \text{otherwise} \end{cases}$$

Then

$$P(Z'_{jk} = 1) = P(Z'_{jk} = 1 | Z_{jk} = 1)P(Z_{jk} = 1) = \frac{p_n}{p_j} \cdot p_j = p_n.$$

So Z'_{jk} can be interpreted as the indicator of bin j being selected under experiment U, and clearly, $P(Z_{jk} \geq Z'_{jk}) = 1$. Observe that $\sum_{k=1}^{m} Z_{jk}$ denotes the number of balls residing in bin j after m trials. So $\min\{1, \sum_{k=1}^{m} Z_{jk}\}\}$ is an indicator function on whether or not bin j is occupied after m trials. Let $\tilde{X} = \min\{m : \sum_{j=1}^{n} \min\{1, \sum_{k=1}^{m} Z_{jk}\} = n\}$ and $\tilde{X}' = \min\{m : \sum_{j=1}^{n} \min\{1, \sum_{k=1}^{m} Z'_{jk}\} = n\}$. Then $P(\tilde{X} \leq \tilde{X}') = 1$. Since $\tilde{X} =_d X$ and $\tilde{X}' =_d X'$, by the definition of stochastic ordering, $X' \geq_{st} X$ so the result follows.

A.2 Proof of Lemma 3

Lemma 3 *We are given n bins with equal probability of being filled with $p_i = p, \ i = 1, ..., n$ and $np \leq 1$. Then the expected time to fill all the bins is H_n/p.*

Proof. Let T denote the time to fill all bins, and define t_i as the time it takes to fill a previously empty bin given i bins are already filled. Then t_i follows a geometric distribution with success probability $(n-i)p$. Therefore,

$$\mathbf{E}[T] = \sum_{i=0}^{n-1} \mathbf{E}[t_i] = H_n/p.$$

A.3 Proof of Theorem 1

Theorem 1*For $\alpha \geq 0$ and $n \geq 1$ bins, the expected number of balls needed to fill all bins have the following bounds.*

$$\frac{1}{c_n^\alpha} \cdot \left(n - \sqrt{n}\right)^\alpha \ln(\sqrt{n}+1) \leq \mathbf{E}[T] \leq \frac{n^\alpha}{c_n^\alpha}(1+\ln n).$$

Proof. To obtain bounds for the harmonic sum, H_n, we use the following inequality

$$\int_0^n \frac{1}{x+1}dx \leq H_n \leq 1 + \int_1^n \frac{1}{x}dx$$
$$\ln(n+1) \leq H_n \leq 1 + \ln n.$$

Using the bounds for H_n, we get the following bounds for $\mathbf{E}[T]$.

$$\frac{(\beta n)^\alpha}{c_n^\alpha} \ln(n - n\beta + 1) \leq \mathbf{E}[T] \leq \frac{n^\alpha}{c_n^\alpha}(1+\ln n) \tag{7}$$

We seek to maximize the lower bound subject to the constraints on β. Only considering the terms involving β, we are interested in maximizing

$$\phi(\beta) = \beta^\alpha \ln(1 + n(1-\beta))$$

over $0 < \beta < 1$. Observe that $\phi(0) = \phi(1) = 0$ and $\phi(\beta) > 0$ when $0 < \beta < 1$. So we can find the maximum by setting the first derivative of ϕ to 0 and solving it.

$$\alpha \ln(1 + n(1-\beta)) = \frac{n}{1 + n(1-\beta))}$$

In order to find a solution for β, we approximate the log function by a linear function.

$$\alpha n(1-\beta) = \frac{n}{1 + n(1-\beta))}$$

Note that even if this approximation is bad for certain values of n or α, the solution for β remains feasible, albeit not optimal. Solving for β, we get

$$\beta = 1 - \frac{\sqrt{\alpha^2 + 4\alpha n} - \alpha}{2\alpha n}$$

but it has the same asymptotic order as the simpler form

$$\beta = 1 - \frac{1}{\sqrt{n}}$$

as n approaches infinity. Plugging this back into $\phi(\beta)$, we get a (near) maximum lower bound. So the resulting bounds for small n is

$$\frac{1}{c_n^\alpha} \cdot \left(n - \sqrt{n}\right)^\alpha \ln(\sqrt{n}+1) \leq \mathbf{E}[T] \leq \frac{n^\alpha}{c_n^\alpha}(1+\ln n).$$

Recall that whenever Lemma 3 is used, in addition to the condition $0 < \beta < 1$, it is also necessary that $(1-\beta)np_{\beta n} \leq 1$ holds.

$$\begin{aligned}(1-\beta)np_{\beta n} &= \frac{c_n^\alpha \sqrt{n}}{(n-\sqrt{n})^\alpha} \\ &< \frac{c_n^\alpha}{\sqrt{n}-1} < 1\end{aligned}$$

is satisfied and the conditions hold.

A.4 Proof of Theorem 2

Theorem 2 *Let* $\mathbf{E}_n[T]$ *denote the expected number of balls needed to fill all* n *bins. For* $\alpha \geq 0$*,*

$$\lim_{n\to\infty} \frac{E_n[T]}{n^\alpha(1+\ln n)/c_n^\alpha} = 1.$$

Proof. Recall the bounds for the expected number of bounds provided by inequality (7). Consider the ratio between the upper and lower bounds

$$\frac{UB}{LB} = \frac{1+\ln n}{\ln(1+n(1-\beta))} \cdot \frac{1}{\beta^\alpha}$$

and let us choose $\beta = 1 - 1/\ln n$. We claim that this ratio converges to 1 as $n \to \infty$, and thus the proposed bounds are tight. Substituting $\beta = 1 - 1/\ln n$,

$$\begin{aligned}1 \leq \frac{UB}{LB} &= \frac{1+\ln n}{\ln\left(\frac{n}{\ln n}+1\right)} \cdot \frac{1}{\left(1-\frac{1}{\ln n}\right)^\alpha} \\ &\leq \frac{1+\ln n}{\ln\left(\frac{n}{\ln n}\right)} \cdot \frac{1}{\left(1-\frac{1}{\ln n}\right)^\alpha} \to 1 \text{ as } n \to \infty\end{aligned}$$

Therefore, asymptotically the bounds are tight and

$$\frac{n^\alpha}{c_n^\alpha}(1+\ln n)$$

is the expected number of balls required to fill all bins when n is large.

We can ensure that the condition $(1-\beta)np_{\beta n} \leq 1$ holds as n approaches infinity since

$$(1-\beta)np_{\beta n} = \frac{c_n^\alpha}{\ln n\left(1-\frac{1}{\ln n}\right)^\alpha n^{\alpha-1}} \to 0.$$

B Proofs of Section 3

B.1 Proof of Theorem 3

Theorem 3 *Let Z be the number of filled bins when T balls are thrown into n bins. When n is large and $\alpha > 1$,*

$$\mathbf{E}[Z] \geq (T+1)^{1/\alpha} - 1.$$

Proof. To be pessimistic, we need to underestimate $\sum_{i=X_T+1}^{n} \frac{c_n^\alpha}{i^\alpha}$.

$$\begin{aligned}
\sum_{i=X_T+1}^{n} \frac{c_n^\alpha}{i^\alpha} &\geq \int_{X_T}^{n} c_n^\alpha (x+1)^{-\alpha} dx \\
&= \frac{c_n^\alpha \left\{ (X_T+1)^{1-\alpha} - (n+1)^{1-\alpha} \right\}}{\alpha - 1} \\
\mathbf{E}[X_{T+1}|X_T] &\geq X_T + \frac{c_n^\alpha \left\{ (X_T+1)^{1-\alpha} - (n+1)^{1-\alpha} \right\}}{\alpha - 1}.
\end{aligned}$$

So

$$\begin{aligned}
&\mathbf{E}[X_{T+1}] \\
&= \mathbf{E}\left[\mathbf{E}[X_{T+1}|X_T]\right] \\
&\geq \mathbf{E}[X_T] + \frac{c_n^\alpha}{\alpha - 1} \mathbf{E}\left[(X_T+1)^{1-\alpha}\right] - \frac{c_n^\alpha}{\alpha - 1}(n+1)^{1-\alpha} \\
&\geq \mathbf{E}[X_T] + \frac{c_n^\alpha}{\alpha - 1} \left\{\mathbf{E}\left[X_T + 1\right]\right\}^{1-\alpha} - \frac{c_n^\alpha}{\alpha - 1}(n+1)^{1-\alpha} \\
&= \mathbf{E}[X_T] + \frac{c_n^\alpha}{\alpha - 1} \left\{\mathbf{E}\left[X_T\right] + 1\right\}^{1-\alpha} - \frac{c_n^\alpha}{\alpha - 1}(n+1)^{1-\alpha} \qquad (8)
\end{aligned}$$

where the last inequality is the result of Jensen's inequality. When n is large, the last term is negligible so we can ignore it. Let $\mathbf{E}[X_T] = f(T)$ and rewriting the inequality, we have

$$f(T+1) - f(T) \geq \frac{c_n^\alpha}{\alpha - 1} \left\{f(T) + 1\right\}^{1-\alpha}.$$

The continuous version of this difference inequality is

$$f'(t)(f(t)+1)^{\alpha-1} = \frac{d}{dt}\left(f(t)+1\right)^\alpha / \alpha \geq \frac{c_n^\alpha}{\alpha - 1}.$$

Integrating both sides and using $f(0) = 0$, we get

$$f(t) \geq \left(\frac{c_n^\alpha \alpha}{\alpha - 1} t + 1\right)^{1/\alpha} - 1.$$

So we have

$$\mathbf{E}[X_T] \geq \left(\frac{c_n^\alpha \alpha}{\alpha - 1} T + 1\right)^{1/\alpha} - 1$$

Therefore, since $\mathbf{E}[X_T]$ is already a lower bound of the true expected number of filled bins, and using the approximation $c_n^\alpha = (\alpha - 1)/\alpha$ when n is large, then

$$(T+1)^{1/\alpha} - 1 \tag{9}$$

is a lower bound to the expected number of filled bins for the original problem.

To be able to ignore the last term of the differential equation (8), $n^{1-\alpha}$ needs to be small. In a realistic setting, it is sufficient that it be smaller than 10^{-5}, which is equivalent to $n > 10^{5/(\alpha-1)}$.

B.2 Proof of Theorem 4

Theorem 4 *For $0 \leq \alpha < 1$, a lower bound on the expected number of filled bins when T balls are thrown is*

$$(n+1)\left(1 - e^{-\frac{(1-\alpha)T}{n+1} - \ln\frac{n+1}{n}}\right) - 1.$$

Proof. The continuous version of (8) is

$$f'(x) \geq \frac{c_n^\alpha}{\alpha - 1}\left[(f(x)+1)^{1-\alpha} - (n+1)^{1-\alpha}\right]. \tag{10}$$

We bound c_n^α using

$$c_n^\alpha \geq \frac{1-\alpha}{n^{1-\alpha} - \alpha} \geq \frac{1-\alpha}{n^{1-\alpha}} \geq \frac{1-\alpha}{(n+1)^{1-\alpha}}$$

and substitute it back in (10) to get

$$f'(x) \geq 1 - \left(\frac{f(x)+1}{n+1}\right)^{1-\alpha}. \tag{11}$$

Using the substitution

$$g(x) = \frac{f(x)+1}{n+1}$$

which is roughly the proportion of filled bins when x balls are thrown, we have

$$g'(x) \geq \frac{1}{n+1}\left(1 - g^{1-\alpha}(x)\right).$$

Observe that $g(x)$ is a monotonically increasing positive function for $x \geq 0$ which converges to 1 as $x \to \infty$. We choose the form

$$g(x) = 1 - e^{-\frac{h(x)}{n+1}}$$

where $h(x) \geq 0$ diverges as $x \to \infty$. Again substituting $g(x)$ with $h(x)$,

$$h'(x) \geq e^{\frac{h(x)}{n+1}} \left[1 - \left(1 - e^{-\frac{h(x)}{n+1}}\right)^{1-\alpha}\right].$$

We claim that $h'(x) \geq 1 - \alpha$ for $x \geq 0$. Since $0 \leq h(x) < \infty$, we can use the substitution

$$t = \frac{h(x)}{n+1}$$

and observe the range of values it can take. We then have

$$h'(x) \geq e^t \left[1 - \left(1 - e^{-t}\right)^{1-\alpha}\right]. \tag{12}$$

The first derivative of the right hand side with respect to t is

$$\begin{aligned}
&e^t \left[(1 - e^{-t})^\alpha - 1\right] + \alpha \\
&\leq e^t \left[\left(1 - \frac{1}{t+1}\right)^\alpha - 1\right] + \alpha \\
&\leq e^t \left[\left(1 - \frac{\alpha}{t+1}\right) - 1\right] + \alpha \\
&= \alpha \left(1 - \frac{e^t}{t+1}\right) \\
&\leq 0.
\end{aligned}$$

So the right hand side of (12) is monotonically decreasing and converges to $1 - \alpha$ as $x \to \infty$. So

$$\begin{aligned}
h'(x) &\geq 1 - \alpha \\
h(x) &\geq (1 - \alpha)x + h(0)
\end{aligned}$$

where since $f(0) = 0$, $h(0) = (n+1)(\ln(n+1) - \ln n)$. Therefore,

$$\begin{aligned}
f(x) &= (n+1)g(x) - 1 \\
&= (n+1)\left(1 - e^{-\frac{h(x)}{n+1}}\right) - 1 \\
&\geq (n+1)\left(1 - e^{-\frac{(1-\alpha)x}{n+1} - \ln\frac{n+1}{n}}\right) - 1,
\end{aligned}$$

which completes the proof.

B.3 Proof of Theorem 5

Theorem 5 *For $\alpha > 0$ and $n \geq 1$, a lower bound on the expected number of filled bins when T balls are thrown is*

$$\left(1 - \left(1 - \frac{1}{T}\right)^T\right)(c_n^\alpha T)^{1/\alpha}.$$

Proof. Recall the formula for the expected number of filled bins $\mathbf{E}[Z]$ when $T \leq n^\alpha / c_n^\alpha$ balls are thrown into n bins:

$$\mathbf{E}[Z] = \sum_{i=1}^{n} \left\{ 1 - \left(1 - \frac{c_n^\alpha}{i^\alpha} \right)^T \right\}.$$

We use the partial sum to find a lower bound.

$$\mathbf{E}[Z] \geq \sum_{i=1}^{k} \left\{ 1 - \left(1 - \frac{c_n^\alpha}{i^\alpha} \right)^T \right\}.$$

where we choose $k = (c_n^\alpha T)^{1/\alpha}$ which forces the condition $T \leq n^\alpha / c_n^\alpha$. Then

$$\begin{aligned}
\mathbf{E}[Z] &\geq \sum_{i=1}^{k} \left\{ 1 - \left(1 - \frac{c_n^\alpha}{i^\alpha} \right)^T \right\} \\
&\geq k - k \left(1 - \frac{c_n^\alpha}{k^\alpha} \right)^T \\
&= k - k \left(1 - \frac{1}{T} \right)^T \\
&= \left(1 - \left(1 - \frac{1}{T} \right)^T \right) (c_n^\alpha T)^{1/\alpha} \qquad (13)
\end{aligned}$$

and the proof is complete.

B.4 Proof of Theorem 6

Theorem 6 *For $n \geq 1$, an upper bound on the expected number of filled bins when T balls are thrown is*

$$\begin{cases} \min\{(4), (15)\} & \alpha = 1 \\ \min\{(4), (14)\} & \alpha \neq 1 \end{cases}$$

Proof. We partition the n bins into 2 metabins where $M_1 = \{1, ..., k-1\}$ and $M_2 = \{k, ..., n\}$. Let $p(k)$ and X_2 denote the probability a ball lands in M_2, and the number of balls that fall in M_2 when T balls are thrown, respectively. Then X_2 follows a binomial distribution $B(T, p(k))$. The number of bins filled in M_1 cannot exceed $k-1$, the cardinality of M_1. Also the number of bins filled in M_2 cannot exceed X_2, the number of balls that land there. Therefore, if Z denotes the random variable that represents the number of filled bins from the same experiment as from which X_2 is measured, i.e., the Z and X_2 reside in the

same probability space, we have the following relation.

$$\begin{aligned} Z &\leq (k-1) + X_2 \text{ (with probability 1)} \\ \mathbf{E}[Z] &\leq (k-1) + \mathbf{E}[X_2] \\ &= k - 1 + Tp(k) \\ &\leq k - 1 + T \cdot \frac{c}{1-\alpha}(n^{1-\alpha} - (k-1)^{1-\alpha}) \\ &:= h(k) \end{aligned}$$

where the last inequality is justified by the following

$$\sum_{i=k}^{n} \frac{c}{i^{\alpha}} \leq \int_{k-1}^{n} cx^{-\alpha} dx.$$

This $h(k)$ is an upper bound for any given k and we seek to find the k that minimizes the upper bound. Observe that $h''(k) = cT\alpha k^{-\alpha-1} \geq 0$ so $h(k)$ is convex in k and has a unique minimum. By solving $h'(k) = 0$ we get

$$k^* = (cT)^{1/\alpha} + 1.$$

Plugging this k back into the expression for $h(k)$ we get an upper bound for the expected number of filled bins.

$$h(k^*) = \frac{\alpha}{\alpha - 1}(cT)^{1/\alpha} - \frac{cT}{\alpha - 1} n^{1-\alpha} \tag{14}$$

Note that $k^* \leq n$, yields

$$T \leq \frac{(n-1)^{\alpha}}{c}$$

which is necessary for $h(k^*)$ to have significance as an upper bound. Otherwise, the maximizing k is when $k = n + 1$ which yields $h(n+1) = n$, a trivial upper bound. Therefore, in such case, we use Cardenas formula for the upper bound. Again, to express the bound in a closed form involving only α, n, and T, we can substitute c by (5). Note that the only condition under which (14) holds is that $\alpha \neq 1$.

When $\alpha = 1$ we have

$$\begin{aligned} p(k) &= 1 - \frac{H_k}{H_n} \\ &\simeq 1 - \frac{\ln k + \gamma}{\ln n + \gamma} \end{aligned}$$

where $\gamma \simeq 0.5772$ is the Euler-Mascheroni constant. Then

$$h(k) = k - 1 + T\left(1 - \frac{\ln k + \gamma}{\ln n + \gamma}\right)$$

is maximized at $k^* = T/(\ln n + \gamma)$ yielding an upper bound

$$h(k^*) = \frac{T}{\ln n + \gamma} - 1 + T\left(1 - \frac{\ln T - \ln(\ln n + \gamma) + \gamma}{\ln n + \gamma}\right). \tag{15}$$

So regardless of the values of α, n, and T, we can use the following upper bound.

$$\mathbf{E}[Z] \leq \begin{cases} \min\{(4), (15)\} & \alpha = 1 \\ \min\{(4), (14)\} & \alpha \neq 1 \end{cases}$$

B.5 Proof of Corollary 2

Corollary 2 *When n is sufficiently large and $\alpha > 1$, the ratio between the lower bound of Theorem 3 and the upper bound of Theorem 6 is less than or equal to $e^{1/e}$.*

Proof. Let LB and UB denote respectively the lower and upper bound provided in the aforementioned theorems. From (5), we have

$$\lim_{n\to\infty} c_n^\alpha = \frac{\alpha - 1}{\alpha}$$

and consequently from (15), we have

$$UB = \left(\frac{\alpha}{\alpha - 1}\right)^{1-1/\alpha} T^{1/\alpha}.$$

Also we have from Theorem 3

$$LB = (T+1)^{1/\alpha} - 1.$$

Considering the coefficients of the dominating terms of UB and LB, the ratio UB/LB is essentially

$$\rho(\alpha) = \left(\frac{\alpha}{\alpha - 1}\right)^{1-1/\alpha}.$$

It can be easily verified that

$$\lim_{\alpha\to 1} \rho(\alpha) = 1$$

and

$$\lim_{\alpha\to\infty} \rho(\alpha) = 1,$$

and that $\rho(\alpha)$ is unimodal and maximized at $\alpha = e/(e-1)$, which yields a maximum value of $e^{1/e} \simeq 1.4447$.